Bias-Variance Analysis of ECOC and Bagging Using Neural Nets

Cemre Zor[1], Terry Windeatt[1] and Berrin Yanikoglu[2]

[1]Center for Vision, Speech and Signal Processing, University of Surrey, UK, GU2 7XH
(c.zor, t.windeatt)@surrey.ac.uk
[2]Sabanci University, Tuzla, Istanbul, Turkey, 34956
berrin@sabanciuniv.edu

Abstract. One of the methods used to evaluate the performance of ensemble classifiers is bias and variance analysis. In this paper, we analyse bagging and ECOC ensembles using bias-variance domain of James [1] and make a comparison with single classifiers, when using Neural Networks (NNs) as base classifiers. As the performance of the ensembles depends on the individual base classifiers, it is important to understand the overall trends when the parameters of the base classifiers, nodes and epochs for NNs, are changed. We show experimentally on 5 artificial and 4 UCI MLR datasets that there are some clear trends in the analysis that should be taken into consideration while designing NN classifier systems.

1 Introduction

Within machine learning research, many techniques have been proposed in order to understand and analyse the success of ensemble classification methods over single-classifier classifications. One of the main approaches considers tightening the generalization error bounds by using the margin concept [6]. Though theoretically interesting, bounds are not usually tight enough to be used in practical design issues. Bias and variance analysis is another method used to show why ensembles work well. In this paper, we try to analyse the success of bagging [22] and Error Correcting Output Coding (ECOC) [4] as ensemble classification techniques, by using Neural Networks (NNs) as the base classifiers within the bias and variance framework of James [1]. As the characteristics of the ensemble depend on the specifications of the base classifiers, having a detailed look at the parameters of the base classifiers within the bias-variance analysis is of importance. Similar work for bagged Support Vector Machines (SVMs) within Domingos' bias-variance framework [7] can be found in [19].

ECOC is an ensemble technique [4], in which multiple base classifiers are trained according to a preset binary *code matrix*. Consider an ECOC matrix C, where a particular element $C_{ij} \epsilon (+1, -1)$ indicates the desired label for class i, to be used in training the base classifier j. The base classifiers are the dichotomizers which carry out the two-class classification tasks for each column of the matrix, according to the input labelling. Each row, called a *codeword*, indicates the desired output for the whole set of base classifiers for the class it is

indicating. During decoding, a given test sample is classified by computing the similarity between the output (hard or soft decisions) of each base classifier and the codeword for each class by using a distance metric, such as the Hamming or the Euclidean distance. The class with the minimum distance is then chosen as the estimated class label. The method can handle incorrect base classification results up to a certain degree. Specifically, if the minimum Hamming distance (HD) between any pair of codewords is d, then up to $\lfloor (d-1)/2 \rfloor$ single bit errors can be corrected.

As for bias and variance analysis, after the initial work of Geman [8] on the regression setting using squared-error loss, others like Breiman [20], Kohavi and Wolpert [10], Dietterich and Kong [9], Friedman [11], Wolpert [23], Heskes [12], Tibshirani [13], Domingos [7] and James [1] have tried to extend the analysis for the classification setting. One of the problems with the above definitions of bias and variance is that most of them are given for specific loss functions such as the zero-one loss, and it is hard to generalize them for all the other loss functions. Usually, new definitions are driven for each loss function. Even if the definitions are proposed to be general, they may fail to satisfy the additive decomposition of the prediction error defined in [8]. The definition of James has advantages over the others as it proposes to construct a scheme which is generalizable to any symmetric loss function. Furthermore, it proposes two more concepts called "systematic effect" and "variance effect" which help assure the additive prediction error decomposition for general loss functions and realize the effects of bias and variance on the prediction error.

Some characteristics of the other definitions which make James' more preferable for us are as follows: 1) Dietterich allows a negative variance and it is possible for the Bayes classifier to have positive bias. 2) Experimentally, the trends of Breiman's bias and variance closely follow James' systematic effect and variance effect ones respectively. However, for each test input pattern, Breiman separates base classifiers into two sets, as biased and unbiased; and considers each test pattern only to have either bias or variance accordingly. 3) Kohavi and Wolpert also assign a nonzero bias to the Bayes classifier but the Bayes error is absorbed within the bias term. Although it helps avoid the need to calculate the Bayes error in real datasets through making unwarranted assumptions, it is not preferable since the bias term becomes too high. 4) The definitions of Tibshirani, Heskes and Breiman are difficult to generalize and extend for the loss functions other than the ones for which they were defined. 5) Friedman proposes that bias and variance do not always need to be additive.

In addition to all these differences, it should also be noted that the characteristics of bias and variance of Domingos' definition are actually close to James', although the decomposition can be considered as being multiplicative [1].

In the literature, attempts have also been made to explore the bias-variance characteristics of ECOC and bagging ensembles. Examples can be found in [1] [9] [20][14][15]. In this paper, a detailed bias-variance analysis of ECOC and bagging ensembles using NNs as base classifiers is given while systematically changing parameters, namely nodes and epochs, based on James' definition.

2 Bias and Variance Analysis of James

James [1] extends the prediction error decomposition, which is initially proposed by Geman et al [8] for squared error under regression setting, for all symmetric loss functions. Therefore, his definition also covers zero-one loss under classification setting, which we use in the experiments.

In his decomposition, the terms "systematic effect" and "variance effect" satisfy the additive decomposition for all symmetric loss functions, and for both real valued and categorical predictors. They actually indicate the effect of bias and variance on the prediction error. For example, a negative variance effect would mean that variance actually helps reduce the prediction error. On the other hand, the "bias" and "variance" terms are defined to show the natural characteristics of the variability and the average distance between the response and the predictor respectively. Therefore, both the meanings and the additive characteristics of the bias and variance concepts of the original setup have been preserved. Following is a summary of the bias-variance derivations of James:

For any symmetric loss function L, where $L(a,b) = L(b,a)$:

$$\begin{aligned} E_{Y,\tilde{Y}}[L(Y,\tilde{Y})] &= E_Y[L(Y,SY)] + E_Y[L(Y,S\tilde{Y}) - L(Y,SY)] \\ &\quad + E_{Y,\tilde{Y}}[L(Y,\tilde{Y}) - L(Y,S\tilde{Y})] \\ prediction\,error &= Var(Y) + SE(\tilde{Y},Y) + VE(\tilde{Y},Y) \end{aligned}$$

where $L(a,b)$ is the loss when b is used in predicting a , Y is the response, $\tilde{Y}$ is the predictor, SE is the systemmatic effect and VE is the variance effect. $SY = argmin_\mu E_Y[L(Y,\mu)]$ and $S\tilde{Y} = argmin_\mu E_Y[L(\tilde{Y},\mu)]$. We see here that prediction error is composed of the variance of the response (irreducible noise), systematic effect and variance effect.

Using the same terminology, the bias and variance for the predictor are defined as follows:

$$\begin{aligned} Bias(\tilde{Y}) &= L(SY, S\tilde{Y}) \\ Var(\tilde{Y}) &= E_{\tilde{Y}}[L(\tilde{Y}, S\tilde{Y})] \end{aligned}$$

When the specific case of classification problems with zero-one loss function is considered, we end up with the following formulations:

$L(a,b) = I(a \neq b)$, $Y \epsilon \{1,2,3..N\}$ for an N class problem, $P_i^Y = P_Y(Y=i)$, $P_i^{\tilde{Y}} = P_{\tilde{Y}}(\tilde{Y}=i)$, $ST = argmin_i E_Y[I(Y \neq i)] = argmax_i P_i^Y$

Therefore,

$$\begin{aligned} Var(Y) &= P_Y(Y \neq SY) = 1 - max_i P_i^Y \\ Var(\tilde{Y}) &= P_{\tilde{Y}}(\tilde{Y} \neq S\tilde{Y}) = 1 - max_i P_i^{\tilde{Y}} \\ Bias(\tilde{Y}) &= I(S\tilde{Y} \neq SY) \end{aligned}$$

$$VE(\tilde{Y},Y) = P(Y \neq \tilde{Y}) - P_Y(Y \neq S\tilde{Y}) = P^Y_{S\tilde{Y}} - \sum_i P^Y_i P^{\tilde{Y}}_i$$

$$SE(\tilde{Y},Y) = P_Y(Y \neq S\tilde{Y}) - P_Y(Y \neq SY) = P^Y_{SY} - P^Y_{S\tilde{Y}}$$

where $I(q)$ is 1 if q is a true argument and 0 otherwise.

3 Experiments

3.1 Experimental Setup

Experiments have been carried out on 5 artificial and 4 UCI MLR [21] datasets. 3 of the artificial datasets are created according to Breiman's description in [20]. Detailed information about the sets can be found in Table 1. The optimization method used in NNs is the Levenberg-Marquart (LM) technique; the level of training (epochs) varies between 2 and 15; and the number of nodes between 2 and 16.

The ECOC matrices are created by randomly assigning binary values to each matrix cell and Hamming Distance is used as the metric in the decoding stage. In the experiments, 3 classification methods are analysed: Single classifier, bagging, and ECOC. In each case, 50 base classifiers are created for bias-variance analysis. Each base classifier is either a single classifier, or an ensemble consisting of 50 bagged classifiers or ECOC matrices of 50 columns.

Experiments have been repeated 10 times for the artificial datasets by using different training & test data,as well as different ECOC matrices in each run; and the results are averaged[1]. The number of training patterns per base classifier is equal to 300; and the number of test patterns is 18000. For the UCI datasets having separate test sets, the analysis has been done just once for the single classifier and bagging settings, and 10 times with different matrices for the ECOC setting. Here, bootstrapping is applied while creating the base classifiers, as it is expected to be a close enough approximation to random & independent data generation from a known underlying distribution [20]. As for the UCI datasets without separate test sets, the *ssCV* cross-validation method of Webb and Conilione [16], which allows the usage of the whole dataset both in training and test stages, has been implemented. In *ssCV*, the shortcomings of the hold-out approach like the usage of small training and test sets; and the lack of inter-training variability control between the successive training sets has been overcome. In our experiments, we set the inter-training variability constant δ to 1/2.

The Bayes error is analytically calculated for the artificial datasets, as the underlying likelihood probability distributions are known. As for the real datasets, the motivation is to find the best optimal classifier parameters giving the lowest error rate possible, through cross-fold validation (CV); and then to use these

[1] On the two class problems, ECOC has not been used, as it would be nothing different than applying bagging. The effect of bootstrapping of bagging would be satisfied by the random initial weights of LM.

Table 1. Summary of the datasets used

	Type	# Training Samples	# Test Samples	# Attributes	# Classes	Bayes Error (%)
TwoNorm [20]	Artificial	300*	18000*	20	2	2.28
ThreeNorm [20]	Artificial	300 *	18000*	20	2	10.83
RingNorm [20]	Artificial	300 *	18000*	20	2	1.51
ArtificalMulti1	Artificial	300*	18000*	2	5	21.76
ArtificalMulti2	Artificial	300 *	18000*	3	9	14.33
Glass Identification	UCI	214	-	10	6	38.66
Dermatology	UCI	358	-	33	6	9.68
Segmentation	UCI	210	2100	19	7	4.21
Yeast	UCI	1484	-	8	10	43.39

*: The training and test samples for the artificial datasets change per each base classifier and per each run respectively.

parameters to construct a classifier which is expected to be close enough to the Bayes classifier. This classifier is then used to calculate the output probabilities per pattern in the dataset. For this, we first find an optimal set of parameters for RBF SVMs by applying 10 fold CV; and then, obtain the underlying probabilities by utilizing the leave-one-out approach. Using the leave-one-out approach instead of training and testing the whole dataset with the found CV parameters helps us avoid overfitting. It is assumed that the underlying distribution stays almost constant for each fold of the leave-one-out procedure.

3.2 Results

In this section, some clear trends found in the analysis are discussed. Although the observations are made using 9 datasets, for brevity reasons we only present a number of representative graphs.

Prediction errors obtained by using bagging and ECOC ensembles are always lower than those of the single classifier; and the reduction in the error is almost always a result of reductions both in variance effect (VE) and in systematic effect (SE). This observation means that the contributions of bias and variance to the prediction error are smaller when ensembles are used (Fig 1, Fig 2). Note that, reductions in VE have greater magnitude, and in two-class problems, the reduction in SE is almost zero (Fig 3). In [20] and [9], bagging and ECOC are also stated to have low variance in the additive error decomposition, and Kong-Dietterich framework [9] also acknowledges that ECOC reduces variance.

The convergence of single classifiers to the optimal prediction error are usually achieved at higher number of epochs than those of bagging; and ECOC ensemble convergence is mostly at even lower epochs than bagging. The prediction errors also turn out in the same descending order: single classifier, bagging and ECOC. The only exceptions to these happen when high number of nodes and epochs are used. Under these circumstances, the VE, SE, and therefore the prediction errors of both ECOC and bagging are similar. However, it should also

be noted that ECOC outperforms bagging in sense of speed due to the fact that it divides multi-class classification problems into binary classification ones.

It is also almost always the case that the prediction error of ECOC converges to its optimum in 2 nodes, whereas a single classifier requires a higher number of nodes. Moreover, for ECOC, the number of epochs at the optimum is also lower than or equal to that of the single classifier. In other words, compared to a single classifier trained with high number of epochs and nodes, an ECOC can yield better results with fewer nodes and epochs. The trend is similar when bagging is considered. It usually stands between the single classifier and ECOC, in sense of accuracy and convergence points.

When the single classifier case is taken into account; we see that VE does not necessarily follow the trend of variance. It happens especially when the number of nodes and epochs is small, that is when the network is relatively weak (Fig 2). In this scenario, the variance decreases while the VE increases. This is actually an expected observation as one would expect having high variance to help hitting the right target class, when the network is relatively less decisive. Ensemble methods do not show this property as much as the single classifier. A possible explanation might be that each base ensemble classifier already makes use of variance coming from the base classifiers it is composed of; and this compensates for the decrease in VE of single classifiers with high variance, in weak networks.

Therefore, having more variance among base ensemble classifiers does not necessarily help having less VE. However, an example of bagging creating negative VE, which clearly states that having variance reduces prediction error; and then going back to positive when variance increases, can be observed on ArtificialMulti2 data when it is processed with 4 node NNs. A similar observation is that although the variance has high values in networks with small number of nodes and epochs, the magnitude of its effect is relatively smaller (Fig 1, Fig 2).

In the above mentioned scenario of VE showing an opposite trend of variance, the bias-variance trade-off can be observed. At the points where the VE increases, SE decreases to reveal an overall decrease in the prediction error. However, these points are not necessarily the optimal points in terms of the prediction error; the optima are mostly where there is both VE and SE reduction (Fig 2). Apart from this case, bias and variance are mostly correlated with SE and VE respectively. This is also pointed out in [1] (Fig 2, Fig 3).

4 Discussion

By analysing bagging, ECOC and single classifiers consisting of NNs through the bias-variance definition of James, we have found some clear trends and relationships that offer hints to be used in classifier design. For multi-class classification problems, the increase in the overall prediction performance obtained with ECOC makes it preferable over the single classifiers. The fact that it converges to the optimum by using smaller number of nodes and epochs is yet another advantage. It also outperforms bagging mostly, while in other cases gives similar results. As for the two-class classification problems, bagging always outperforms

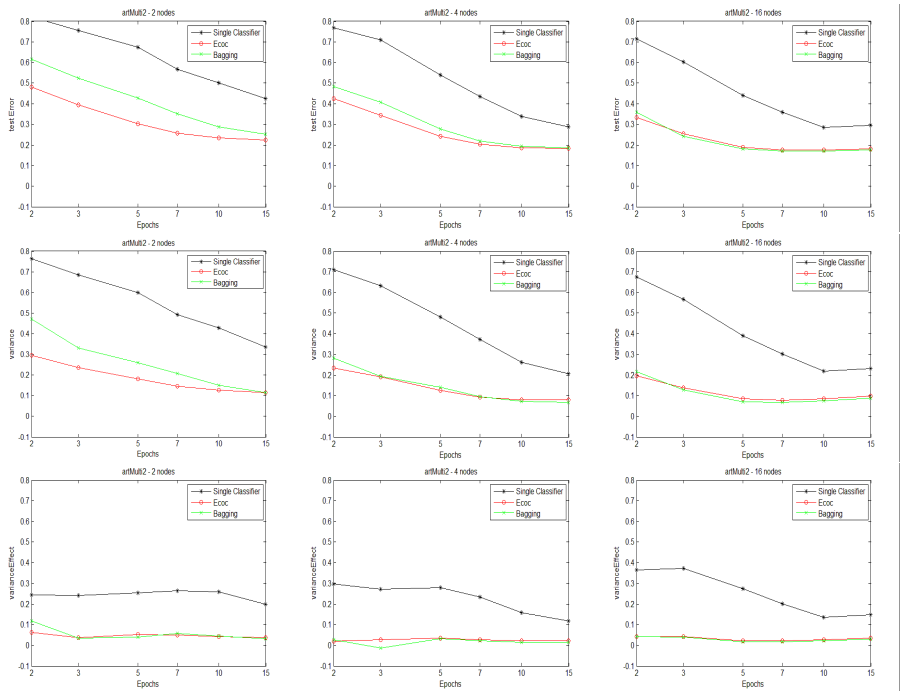

Fig. 1. Bias Variance Analysis for ArtificalMulti2 data. First Row: Overall prediction error. Second Row: Variance. Third Row: Variance effect. First Column: For 2 Nodes. Second Column: For 4 Nodes. Third Column: For 16 Nodes. Black lines indicate the results for single classifier, red for ECOC and green for bagging

the single classifier; and the optimum number of nodes and epochs is relatively smaller.

The increase in the performance of bagging and ECOC is a result of the decrease in both variance effect and systematic effect, although the reductions in the magnitude of the variance effect are bigger. Also, when the NNs are weak, that is when they have been trained with few nodes and epochs, we see that the trends of variance and variance effect might be in opposite directions in the single classifier case. This implies that having high variance might help improve the classification performance in weak networks when single classifiers are used. However, they are still outperformed by ensembles, which have even lower variance effects.

As for further possible advantages of ensembles, the fact that they are expected to avoid overfitting might be shown by using more powerful NNs with higher number of nodes, or other classifiers such as SVMs that are more prone to overfitting. Future work is also aimed at understanding and analysing the

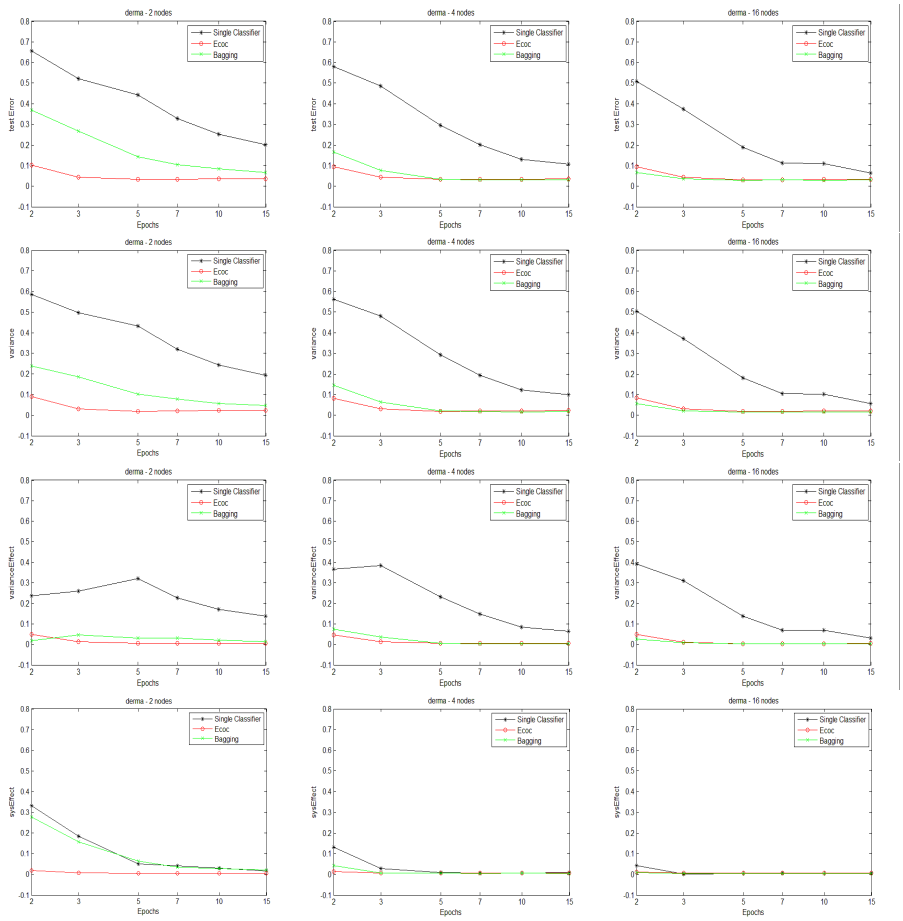

Fig. 2. Bias Variance Analysis for Dermatology data. First Row: Overall prediction error. Second Row: Variance. Third Row: Variance effect. Fourth Row: Systematic effect. First Column: For 2 Nodes. Second Column: For 4 Nodes. Third Column: For 16 Nodes. Black lines indicate the results for single classifier, red for ECOC and green for bagging

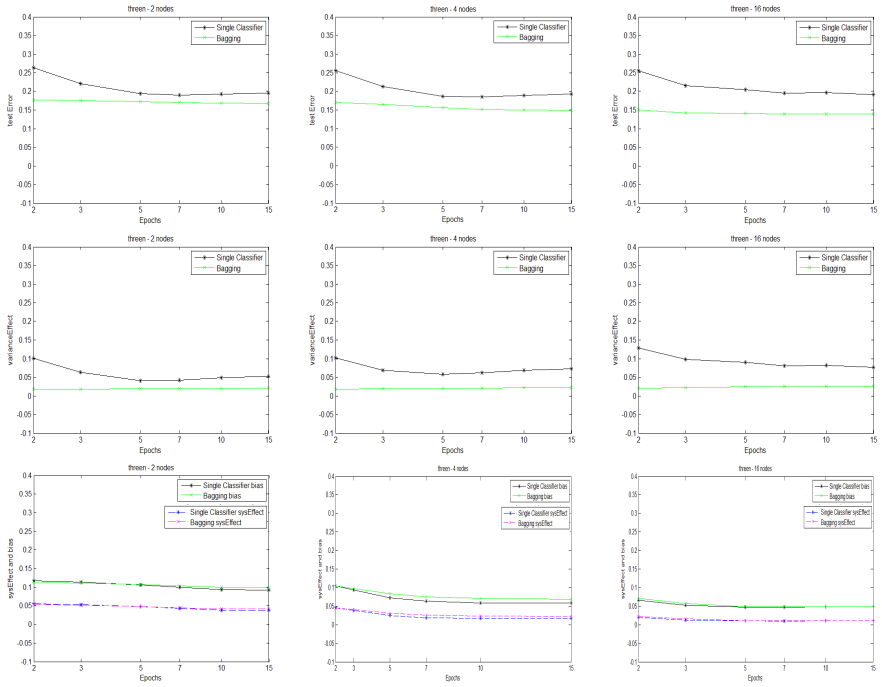

Fig. 3. Bias Variance Analysis for ThreeNorm data. First Row: Overall prediction error. Second Row: Variance effect. Third Row: Systematic effect and Bias. First Column: For 2 Nodes. Second Column: For 4 Nodes. Third Column: For 16 Nodes. Black & blue lines indicate the results for single classifier (bias and systematic effect) and green & magenta for bagging

bias-variance domain within some mathematical frameworks such as [17] [18] and using the information in the design of ECOC matrices.

References

1. James, G.: Variance and Bias for General Loss Functions, Machine Learning, 51(2), 115–135 (2003)
2. Dietterich, T.G., Bakiri, G.: Solving Multi-class Learning Problems via Error-Correcting Output Codes. J. Artificial Intelligence Research 2. 263–286 (1995)
3. Allwein, E., Schapire, R., Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. JMLR 1. 113–141 (2002)
4. Dietterich, T.G., Bakiri, G.: Solving Multi-class Learning Problems via Error-Correcting Output Codes. J. Artificial Intelligence Research 2. 263–286 (1995)
5. Allwein, E., Schapire, R., Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. JMLR 1. 113–141 (2002)
6. Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S.: Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 26(5):1651–1686 (1998)

7. Domingos. P.: A Unified Bias-Variance Decomposition for Zero-One and Squared Loss. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence, pp. 564–569 (2000)
8. Geman, S., Bienenstock, E., Doursat R.: Neural networks and the bias/variance dilemma, Neural Comput., vol. 4, no. 1, pp. 1-58 (1992)
9. Kong, E. B., Dietterich, T. G.: Error-correcting Output Coding Corrects Bias and Variance. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 313-321 (1995)
10. Kohavi, R., & Wolpert, D. H.: Bias plus variance decomposition for zero-one loss functions. In: Proceedings Thirteenth International Conference on Machine Learning, pp.275- 283 (1996)
11. Friedman, J. H.: On bias, variance, 0/1 loss and the curse of dimensionality. Data Mining and Knowledge Discovery, 1, pp. 55–77 (1997)
12. Heskes, T.: Bias/Variance Decompostion for Likelihood-Based Estimators. Neural Computation, 10, pp. 1425–1433 (1998)
13. Tibshirani, R.: Bias, variance and prediction error for classification rules. Technical Report, University of Toronto, Toronto, Canada (1996)
14. Smith, R. S., Windeatt, T.: The Bias Variance Trade-Off in Bootstrapped Error Correcting Output Code Ensembles. MCS, pp.1–10 (2009)
15. Domingos, P.: Why does bagging work? A Bayesian account and its implications .Proceedings of the 3rd International Conf. on Knowledge Discovery and Data Mining, pp. 155–158 (1997)
16. Webb, G.I., Conilione, P.: Estimating bias and variance from data. Technical Report, (2005)
17. Tumer, K., Ghosh, J.: Error correlation and error reduction in ensemble classifiers. Connection Science 8 (3-4), pp. 385–403 (1996)
18. Tumer, K., Ghosh, J.: Analysis of decision boundaries in linearly combined neural classifiers. Pattern Recognition, 29(2), pp. 341–348 (1996)
19. Valentini, G., Dietterich, T.: Bias–variance analysis of Support Vector Machines for the development of SVM-based ensemble methods. Journal of Machine Learning Research, vol. 5, pp. 725-775 (2004)
20. Breiman L.: Arcing classifiers. The Annals of Statistics, 26(3), 801-849 (1998)
21. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html. School of Information and Computer Science, University of California, Irvine, CA (2007)
22. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the 13th ICML, pp. 148–156 (1996)
23. Wolpert, D. H.: On bias plus variance. Neural Computation. 9, pp. 1211-1244, (1996)